Exploring the Unknown – Collecting 1000 speakers over the Internet for the Ph@ttSessionz Database of Adolescent Speakers

Christoph Draxler

BAS Bavarian Archive for Speech Signals
Institut für Phonetik und Sprachliche Kommunikation
Ludwig-Maximilians-Universität München, Munich, Germany
draxler@phonetik.uni-muenchen.de

Abstract

The Ph@ttSessionz project will create a database of 1000 adolescent German speakers. The project employs a novel approach to collecting speech data: recordings are being performed via the WWW in more than 35 schools in Germany, and the data is immediately transferred to the BAS server in Munich. Using this approach, geographically distributed recordings in high bandwidth quality can be performed efficiently and reliably. The paper presents the infrastructure developed at BAS for WWW-based speech recordings, it discusses the strategies employed to get schools to participate in the project, and it presents preliminary analyses of the speech database.

Index Terms: speech database collection, web-based recording, transcription, client-server system, adolescent speech.

1. Introduction

The development of speech technology requires large speech databases for training and testing. Ideally, these databases contain speech material from the typical users of a given technology, and from the application domain of the technology.

The last years have seen the collection of a large number of speech databases for technology development: for the telephone network (e.g. Switchboard [1], Macrophone [2], SpeechDat [3]), high bandwidth recordings (e.g. Speecon [4]), mobile environments (e.g. SpeechDat-Car [5]), information and dialog systems (e.g. Verbmobil [6], SmartKom [7]), etc.

These databases show the well-known trade-off between the size of a database and the signal quality: very large databases with more than 500 speakers and a good geographical coverage exist only for telephone speech; high bandwidth databases typically have less than 300 speakers and must thus compromise on the demographic or recording environment criteria (e.g. in SpeechDat-Car and SPEECON, every speaker was recorded twice, in different environments). Clearly, this is subobtimal, and a new approach to collecting speech data is needed.

At BAS we have developed a web application for performing speech recordings via the WWW in a client server architecture. The server provides the data storage and administration, the clients perform the recordings. This approach has a number of important advantages: high quality recordings can be distributed geographically, they can be performed in parallel, and all data is transferred to the server immediately. On the server, recording sessions can be monitored online, and prompt scripts can be adapted flexibly to correct any deviations early.

Ph@ttSessionz is a pilot study for web-based recordings. An account of the technical problems that had to be solved in Ph@ttSessionz was given in [8]. This paper focuses on the recruitment of schools and speakers, and presents some preliminary results.

2. Ph@ttSessionz database

The Ph@ttSessionz database is a superset of the RVG-1 [9] and the SpeechDat fixed telephone network database [10] (table 1).

Table 1: Ph@ttSessionz database contents

type	code	count
isolated digits, "zwo"	01-11	10
numbers 11-100	12-30	19
PC command phrases	31-42	12
phonetically rich sentences	43-72	30
telephone numbers with area code	73-75	3
6- and 7-digit telephone numbers	76-85	10
mobile phone keys (digits, $*$, #)	B1-B3	3
credit card numbers (16 formatted digits)	C1-C3	3
PIN codes (6 digits)	C4-C6	3
date expressions	D1-D3	3
spellings (artificial sequences, names)	L0-L9	10
directory assistance names	O1-O9	9
time expressions	T1-T3	3
spontaneous responses	X1-X5	5
narrative speech	Y1-Y2	2
test items	Z0-Z4	4

2.1. Speaker demographics

The speakers in Ph@ttSessionz are between 13 and 18 years old. Voices of adolescent speakers are particularly interesting because they exhibit a wide range of variability, and because no publicly available database for speech technology development exists with speakers of this age.

The following demographic data is collected for every speaker:

- age at the time of recording, sex, size, weight,
- smoking habits, presence of dental braces or piercings,

- native language of the speaker and his parents,
- federal state where the speaker entered school.

The distribution of sex should be balanced with a tolerance of 5%.

2.2. Regional coverage

The dialect regions identified by [12] are used in Ph@ttSessionz. In each of these regions, we attempt to collect at least 50 speakers. Within a region, we will collect speech in at least two locations (Fig. 1).

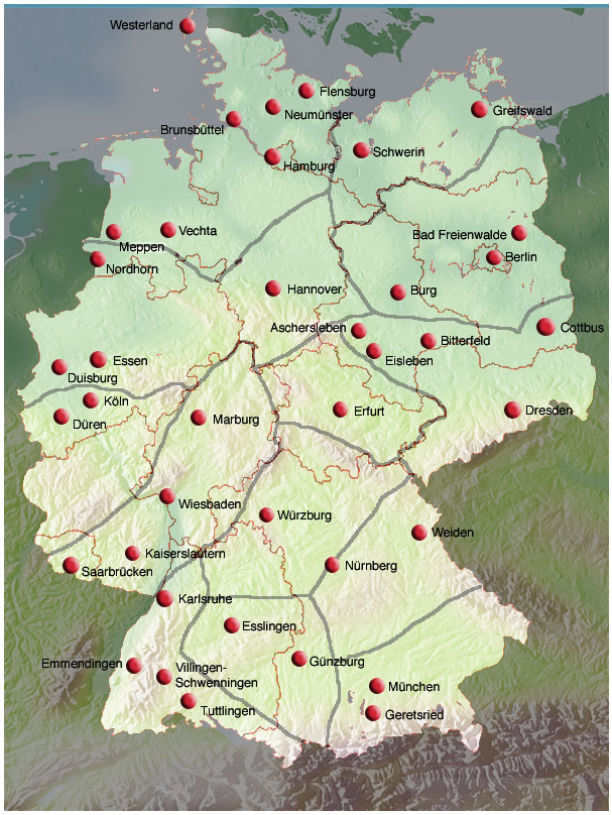

Figure 1: Ph@ttSessions recording locations, federal states and dialect regions according to [12]

2.3. Signal quality

Recordings are performed in public high schools. These schools generally are connected to the Internet with a high-speech DSL connection with a minimum speed of 2048 kBit/s downlink and 256 kBit/s uplink.

To guarantee a consistent signal quality, a standard recording equipment is used at all sites. It consists of a Beyerdynamic opus 54 close talk microphone, an Audio Technica 3031 desktop microphone, and an M-Audio Mobile-Pre USB microphone amplifier and analog/digital converter. The sample rate is 22.05 kHz, with a quantization of 16 bit linear. The signal data rate thus is 705.6 kBit/s.

3. Software and infrastructure

At BAS we have set up a web site for the Ph@ttSessionz speech data collection: `www.phonetik.uni-muenchen.de/phatt`. This web site serves as a portal to all Ph@ttSessionz-related services, and it is divided into six major areas: general information, hardware setup instructions, speech recordings, speech annotation, data collection statistics and documentation. It distinguishes four classes of users: general public, experimenters, transcribers and administrators. All users except the general public have to log in to access their services.

The general information pages contain an introductory text to the project, and photos, a demo video, press articles and sample recordings. The instructions pages describe the hardware and how to set it up at a local site. The recordings pages allow creating new recording sessions: here the experimenters enter the speaker data and start the recording session, each with its own customized prompt sheet. The transcription pages start the transcription software, and the statistics pages give a short report on the current status of the data collection (list of participating schools, timestamps of most recent 20 recording sessions, etc.).

The web pages are implemented as a web application, i.e. the server stores the data and the application logic, and clients connect to the services. The clients are standard web browsers for the portal, and the standalone web-aware applications SpeechRecorder [13] for the recording and WebTranscribe [14] for the transcription.

The web application is implemented using Java Server Pages technology on a Linux PC running the Apache Tomcat web and application server in conjunction with a PostgreSQL relational database system.

4. Recruitment

The Ph@ttSessionz speech data collection is very different from any previous data collection by BAS. Recording adolescent speakers under the legal age (18 years in Germany) raises privacy and legal issues – parents need to be informed and they have to express their consent. Furthermore, the recordings take place in many geographically distributed locations in parallel – these locations had to be found and then supported during the recordings.

After some discussion on how to best reach the target speaker population we agreed to perform the recordings in public schools, and to delegate the task of recruiting speakers to these schools. As an incentive to participate we offered each school 200 Euro for a complete set of recordings. The school was free to use the money in any way it wanted. Furthermore, all schools are listed as contributors on the project web site.

In a first round of the recruiting campaign we selected candidate schools according to the following criteria.

- The school should be a high school (*Gymnasium*) because of their good communication infrastructure,
- it should be in a larger city characteristic for a dialect region, and
- it should be large enough to recruit and record 50 speakers.

We set up a team of three people to call one school after the other to ask for participation. In the beginning, these telephone interviews were quite spontaneous and ad hoc, and they were not very effective. We then worked out a guideline for telephone interviews, which helped a) to standardize communication with the schools,

b) contained the most important information on the project in clear and precise formulations, and c) allowed us to determine quickly whether a school was interested or not.

We quickly realized that asking for 50 speakers was too much – after we reduced the number to 30 speakers, the first schools agreed to take part in the project.

With fewer speakers per school – and with longer recording periods than anticipated – we had to recruit additional schools.

We thus turned to high schools in smaller cities. Contrary to our original belief, these schools are actually larger than those of larger cities, which made recruiting speakers a bit easier. Furthermore, most of these schools have their own system administrator, which facilitated operating system and Java updates which were necessary to run the SpeechRecorder software.

A very successful scheme was to contact the local press in the city of the school. We informed the press about the project, and invited journalists to participate in a recording session. We put press statements, sample articles, photos from recordings in other schools, etc. on our web site, that the newspapers could use.

Currently we have sufficiently many schools in the queue to reach 1000 speakers by the end of June 2006. However, we are still recruiting schools in those dialect regions where only a few speakers have been recorded, e.g. Ruhrgebiet and the Mosel region (regions NW and RP in Table 2).

5. Database statistics

By April 3, 2006, we have recorded 714 speakers with more than 125 spoken items each.

5.1. Speakers

348 (48.74 %) speakers are female, 366 are male (51.26 %). The dialect distribution is given in Table 2.

Table 2: Federal state where speakers entered school

federal state	count	% recordings	% population
BB	64	8.96	3.11
BE	26	3.64	4.11
BW	67	9.38	12.99
BY	82	11.48	15.08
HB	1	0.14	0.80
HE	48	6.72	7.38
HH	0		2.10
MV	36	5.04	2.08
NI	69	9.66	9.70
NW	69	9.66	21.91
RP	8	1.12	4.92
SH	77	10.78	3.43
SL	28	3.92	1.28
SN	37	5.18	5.21
ST	77	10.78	3.02
TH	14	1.96	2.85
other	11	1.54	

The federal state is an information that speakers can easily provide. Although the federal states do not match the dialect regions shown in Figure 1 exactly, they serve as a good starting point for the accent classification.

Of the 714 speakers, 614 (85.99%) declared they were non-smokers, 100 (14.01%) smokers. 637 (89.22%) did not wear dental braces, 77 (10.78%) did. 709 (99.30%) speakers have no piercing in their tongue or lips, 5 (0.70%) have them.

679 (95.1 %) speakers were native German speakers, 655 (91.73 %) said German was their mother's native language, 645 (90.33 %) their father's.

5.2. Recording sessions

Figure 2 shows the number of recording sessions per month. The low figures for August 2005 are due to the holiday season, the higher numbers after August 2005 reflect the fact that the number of recording equipment sets was increased from 5 to 8.

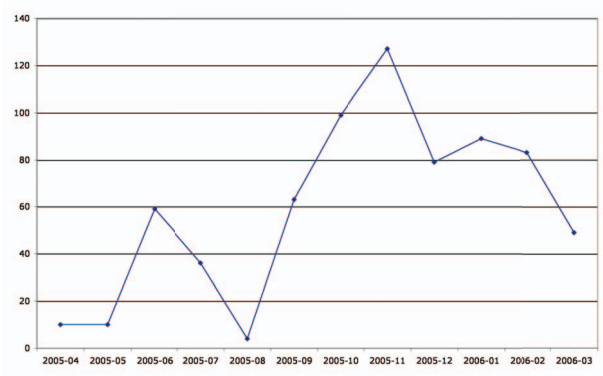

Figure 2: Recording sessions per month

Especially in the beginning of the project schools complained about the duration of recording sessions, which was longer than planned, and was a result of low data transfer. With an improved software, this problem was less severe, and more than 64.7 % of the sessions took less than 40 minutes (Fig. 3).

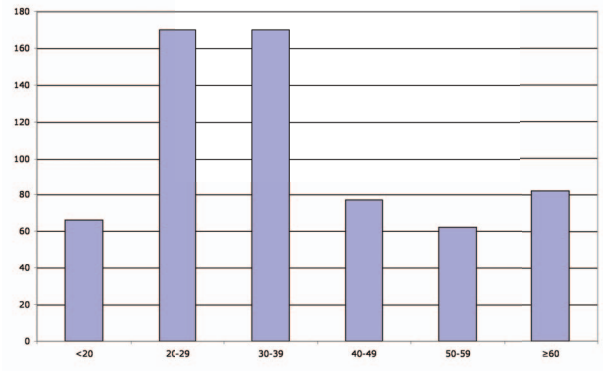

Figure 3: Duration of recording sessions in minutes

Originally we had planned to allocate three weeks for the recording period at a school. Most schools took longer – setting up the equipment, performing the required tests, etc. added substantial delays. However, in the course of the project we did find a moderate reduction in the duration of the recording periods (Fig-

ure 4). It is not clear whether this speedup is due to the improved software which allows schools to define tighter recording schedules, or the increased number of phone calls from our recruiters whenever we found a school to be behind schedule.

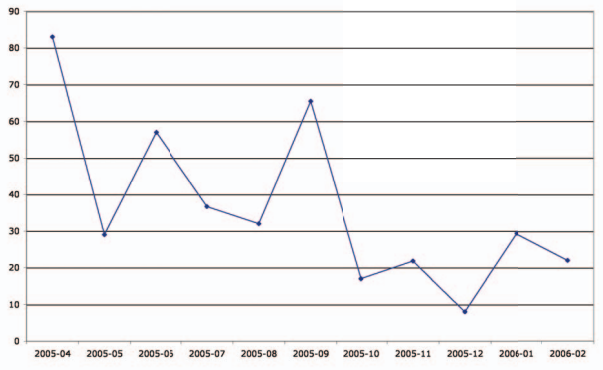

Figure 4: Development of the average recording period duration in the schools (in days) over the project duration

5.3. Annotation

By April 3, 2006, 16978 items have been transcribed (18.82%), covering slightly more than 12 hours of speech. The annotation scheme is an orthographic transcription according to the SpeechDat guidelines [10].

Table 3 gives preliminary results for the annotation of the Ph@ttSessionz recordings. Note that for the read items, i.e. sentences, spellings, digit strings and isolated digits, annotation times of more than 10 minutes were discarded from these calculations because they are artefacts: the annotators did not close the annotation session when they left their terminal for a break. 1782

type	code	segment length	annotation time
single digit	01-10	00:00.73	00:09.70
mobile phone keys	B1-B3	00:05.36	00:21.34
credit card number	C1-C3	00:07.25	00:28.49
spelling	L0-L9	00:04.69	00:18.54
read sentences	43-72	00:02.74	00:17.93
spontaneous response	Y1-Y3	00:11.35	02:26.77

Table 3: Average segment lengths and annotation times for a subset of the Ph@ttSessionz speech database (format *min:sec.msec*)

(10.5 %) transcriptions contain markers for speaker or non-speaker noise, signal truncations or mispronunciations.

6. Conclusions and Outlook

With the Ph@ttSessionz recordings we have shown that large-scale microphone quality speech databases with geographically distributed recording locations are feasible via the WWW. The technology is now mature, and recruitment and organizational procedures have been established to minimize administrative overhead.

The development of a coherent web application has shown to be a major advance in the collection, transcription and validation of speech databases. We plan to extend and generalize the Ph@ttSessionz web application into a full speech database collection framework to facilitate future speech data collections.

7. Acknowledgements

The author would like to thank all schools for their support and the efforts they put into the recordings. Furthermore, I thank Klaus Jänsch for his excellent programming, the BITS team for the recruitment and support work, and the many students for transcribing the Ph@ttSessionz recordings.

The project is funded by BMBF under grant no. 01IVB01.

8. References

[1] Godfrey J., Holliman E., McDaniel J., Switchboard: Telephone Speech Corpus for research and development, Proc. of ICASSP, 1992, San Francisco.

[2] Bernstein J., Taussig K., Godfrey J., Macrophone: An American Speech Corpus for the PolyPhone project, Proc. of ICASSP, 1994.

[3] Höge H., et al., "SpeechDat Multilingual Speech Databases for Teleservices: Across the Finish Line", Proc. of Eurospeech, 1999, Budapest.

[4] Siemund R., Höge H., Kunzmann S., Marasek K., SPEECON – Speech Data for Consumer Devices, Proc. of LREC, 2000, Athens.

[5] van den Heuvel H., Bonafonte A., Boudy J., Dufour S., Lockwood Ph., Grudszus R., SpeechDat-Car: Towards a Collection of Speech Databases for Automotive Environments, Proc. of ICASSP, 1999.

[6] Burger S., Weilhammer K., Schiel F., Tillmann H., Verbmobil Data Collection and Annotation. in: Verbmobil: Foundations of Speech-to-Speech Translation (Ed. Wahlster, W.); Springer; Berlin/Heidelberg.

[7] Schiel F., Steininger S., Türk U., The SmartKom Multimodal Corpus at BAS. Proc of LREC, 2002, Gran Canaria

[8] Draxler Chr., Jänsch K., Speech Recordings in Public Schools in Germany - the Perfect Show Case for Web-based Recordings and Annotation, Proc. of LREC, 2006, Genova.

[9] Burger S., Schiel F., "RVG-1 – A Database for Regional Variants of Contemporary German", Proc. of LREC, 1998, Granada.

[10] Winski R., Definition of Corpus, Scripts, and Standards for Fixed Networks, SpeechDat Report LE2-4001-SD1.1.1, 1997

[11] Bavarian Archive for Speech Signals, `www.bas.uni-muenchen.de`.

[12] Hollmach, U. "Untersuchungen zur Kodifizierung der Standardaussprache in Deutschland", Habilitationsschrift, Universität Halle, 2003.

[13] Draxler Chr., Jänsch K., "SpeechRecorder – A Universal Platform Independent Multi-Channel Audio Recording Software", Proc. of LREC, 2004, Lisbon.

[14] Draxler Chr., "WebTranscribe – An Extensible Web-Based Speech Annotation Framework", Proc. of TSD 2005, LNCS, Springer Verlag, 2005, Berlin.